Performance Evaluation of a Simple Deep Neural Network System for Auditory Attention Decoding

by

Jack Magann

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science in Engineering

Baltimore, Maryland

May 2020

Abstract

Recent years have seen an increase in the variety of methods used to perform Auditory Attention Decoding (AAD). Current high performing methods for auditory attention decoding rely on large training sets and lack comparable standards with one another largely due to the variability in the training data used. Simple standards between these models could help researchers better interpret performance and direct the progression of work to a model that performs the best. Here the performance of a Deep Neural Network (DNN) architecture for AAD proposed by (Cicarelli et al.,2019) is evaluated on a new, smaller set of training data collected in (Fuglsang et al.,2017). The network is shown to successfully achieve learning behavior when presented with the reduction of training data. Limiting the number of listeners used for training based on the output average loss curve resulted in comparable decoding accuracy. Further metrics show the benefit of an analysis of the relevance of listeners used for training of the network. The consistent performance of the network given the reduction in the provided training data shows how the simple DNN is a robust method for performing AAD. It also allows us to properly compare the performance of the DNN with the linear method in (Fuglsang et al.,2017).

Contents

List of Figures

Chapter 1

Introduction

1.1 Problem Statement

In a single day a person is exposed to a plethora of sounds from a variety of sources, each containing a dense amount of information regarding our surroundings and interpersonal communication. Being able to distinguish and organize this constant and complex auditory input is one of the most impressive capabilities of the human brain. In particular, our brain can constantly adapt and change almost instantaneously to determine which of these sources is the most important and should be focused on. It's even just by nature that one can simultaneously track various auditory stimuli in an environment. When speech streams are among the mixed audio input it is instinctive for people to tune their attention to that. A key aspect of transferring information through language is the need of listeners to correctly hear enough content. However, when multiple speech signals are present most find it difficult to maintain accurate information retention though are highly aware of the other speakers' presence and location. Current research is highly interested in how the brain is able to accurately process such complex audio scenes (Han et al., 2019, M¨uller et al., 2020, Vandecappelle et al,2020, Taillez et al., 2020). Particularly, efforts are being made to understand how the brain steers attention in the presence of multiple sources of human speech. The commonly studied task is a person distinguishing a speech source from others while being able to retain the information from that specified source to an acceptable degree. This thesis explores

the recreation of nonlinear decoding techniques in replicating a simplified version of an auditory attention decoding (AAD) task and its robustness to a newly introduced data set.

The classical motivation for mimicking auditory attention is described by the Cocktail Party Problem (Cherry et al., 1953). This problem refers to a person's ability to attend to one speaker when multiple are present in a complex auditory environment, such as at a crowded cocktail party. Though it is a basic communication task that is performed unconsciously throughout every day and by humans of all ages it is an intrinsically difficult problem when it comes to replicating how humans solve it. Selective attention is an intrinsic part to everyday life and is has been shown that the ability to distinguish the unique physical features of speech from background noise is learned from a young age (Plude et al., 1994). Specific auditory queues, such as a person's name, are shown to also be coded to grab our attention at a very young age (Newman, 2005) and require much less perceptual information to do so (Driver, 2001). As adults humans have a much stronger comprehension of human speech, however in a multi-speaker environment maintaining attention suffers when the information in the speech streams are determined to be equally important (Plude et al., 1994). Beyond the auditory aspect multiple types of sensory queues are used in these situations to determine where ones' attention should be. Research has shown that visual salience and auditory attention are likely controlled by similar neural mechanisms (Shinn-Cunningham, 2008). In the scope of this work only the auditory aspect of determining attention in the Cocktail Party Problem is of interest.

1.2 Mechanisms of the Brain

1.2.1 Auditory Pathway

For decoding all of these complex auditory signals the brain has an equally complex system, referred to as the auditory pathway. This sensory system helps in translating and processing all the audio information from the environment into electrical signals in the auditory cortex. This system starts with the physical input of sound being received by the outer ear, consisting of the ear canal and ear drum, and then compressed by the middle ear, an interconnected structure of three bones. This

is then transferred to the inner ear where the cochlea applies nonlinear processing to convert the input audio to a series of electrical spikes in the auditory nerve. These spikes are then processed by neurons in the brainstem, midbrain auditory cortex, and higher cortical areas. It has been shown that along the auditory pathway many of the physical characteristics of each individual stimulus are processed whether or not the listener is choosing to attend to it (Naatanen 90'). Such characteristics include frequency content (Schreiner et al., 2000), localization information (Brugge et al., 2008), and physical envelope shape (Ding et al., 2012). This breakdown of information is critical in our brain's ability to accurately segregate the multiple audio streams. In each level of the auditory pathway a tonotopic organization scheme exists (Humphries et. el). In a tonotopic organization neurons are specifically organized by their response to different frequencies. The arrangement of neurons mirrors the distribution of receptors in the cochlea's basilar membrane. This structure extends from neurons that better respond to high frequencies to neurons that better respond to low frequencies. It has been hypothesized that it is through a combination of the extracted characteristics in a consistent organization that the brain is able to decode the surrounding auditory environment and construct appropriate auditory objects to represent the individual input streams (Shihab et al., 2011).

1.2.2 Attention

Another mechanism that affects the brain's ability to perform source separation, information retention, and source identification is attention. Though the brain does catalog all the incoming sources it is through active and passive attention mechanisms that allow the brain to segregate the incoming auditory streams. Selective auditory attention allows listeners to suppress interfering audio streams and focus on the desired, relevant stream (Bregman and Ahad, 1996). Furthermore is has been shown how selective auditory attention is associated with the entrainment of neural oscillations (Zion Golumbic et al., 2012). This entrainment pertains to how low-frequency brain oscillations synchronize to the temporal regularities of input auditory stimuli. Selective attention causes these neural oscillations to modulate their phase to align high excitability phases to critical events in the chosen attended audio input (Zoefel and VanRullen, 2016). This type of neural filtering has show to have a cross-sensory effect since a similar behavior is seen in such cortical regions

as the visual cortex (Fiebelkorn et al. 2013). The unattended audio streams should then provide fewer similarities with neural oscillations.

Many applications of selective source enhancement have tried to reproduce the brains highly accurate process for the task. Of particular interest is how the brain determines attention and uses that to control source separation. There are two well-established models for describing auditory attention. First, in a bottom-up attention model the mechanism that determines attention is driven by the content of the stimulus. This would imply that a new sound or speech source with certain qualities would be perceived as more salient, or noticeable, to the listener. An example would be a loud, unexpected bang or one's name being called. It has been shown that this basic drive to pick out deviancy in dynamic auditory scenes plays a role in how we form the incoming auditory objects (Kaya and Elhilali, 2014). Second, a top-down attention model is task-dependent where attention is selectively shifted towards what the listener wants to attend. In this model the listener actively uses prior knowledge, such as a speaker's gender or a learned association between a fire alarm and danger, to focus attention on a single input of the overall auditory scene. In using this model the brain is implied to be using the momentary memory of the audio scene to update the attended audio. It is widely acknowledged that these two models work in sync to help shape a listener's understanding of an auditory scene, but the extent that each plays is still largely not understood.

1.3 Auditory Attention Decoding

In Auditory Attention Decoding (AAD) of speech many studies focus on applying a top-down model for determining the attended input speech stream of a listener (Fuglesang et al., 2017,Kalinli ,2008, Ciccarelli et al., 2019). Naturally the brain can quickly separate and determine the desired attended audio stream due to the processing in the auditory cortex breaking down the auditory scene. Once the audio stream is identified the brain is able to boost that signal, while suppressing all other unattended speech and extraneous noise. A key aspect of this model is determining how the individual characteristics of the multiple input audio streams are received and separated by the listener to pick out the desired audio stream. To this end simple comparison techniques use reference audio, representing the desired attended audio, to compare incoming audio streams to

identify the target stream and enhance it (Kalinli ,2008). This requires access to the individual clean signals or a functional source separation algorithm and knowledge of what you are looking for. Applying this technique is ineffective for performing a real-world AAD task as it is impossible to have access to the attended audio in real time prior to it reaching the listener.

Recent years have brought researchers closer in replicating the brain's ability to identify attended audio stream in that speaker identification, source separation, and machine learning techniques are being integrated into how the attended speaker is being identified. A natural progression is a system that is trained to recognize a person's voice through provided samples that allow the system to store the characteristics as comparison criteria for future detected input audio streams. These techniques have been very developmental to the integration of technology in our everyday lives, but again suffer due to an inability to adapt in real-time to new situations where a large set of training data is not available. Situations where the incoming audio is not known prior to receiving it are called blind source-separation tasks. Since the brain is naturally attuned to solving such tasks it is clear that relevant AAD models must have this functionality and that the brain's real-time response to an auditory scene can act as a tool in this.

For a blind source-separation task to successfully model the complex inner working of the brain there are many factors to consider. To determine attention the goal is to decode the brain's response to the training input audio so that one can predict, based on the auditory cortex response, which one the listener is attending to. The response of the cerebral cortex to speech has been shown to be correlated with the envelopes of both the unattended and attended audio (Ding 2011, Mesgarani et el., 2012), especially in the delta (1-4 Hz) and theta (4-8 Hz) bands (Ding et al., 2012). This has led researchers to propose the use of electroencephalography (EEG) (O'sullivan, 2014), magnetoencephalography (MEG) (Akram, 2017), and electrocortiography (ECoG) (Dijkstra, 2015) recordings to reconstruct the envelope of the attended audio speech signal. Of key interest with these neural recordings is the measurement of auditory event-related potentials (ERP) or event-related fields (ERF), which are the neural responses resulting directly from a specific auditory sensory event. Typical AAD implementations then compare each individual audio stream's acoustical properties to the neural data and the one with the highest similarity is determined to be the target of attention.

Through training of multiple methods this can be effective in decoding where a listener desires their attention to go. In most applications this would result in an enhancement to the desired input audio stream. One commercially popular method for replicating this selective enhancement is beamforming where increased sensitivity is put on audio coming from a specific direction around the speaker. Basic implementations determine a particular direction of enhancement by either pre-setting the direction or choosing the direction of the loudest input audio. These methods are easily open to a lot of error, such as when background noise is louder than the speech stream of interest or the pre-set direction does not match the location of the speech stream. In accurately reproducing the brain's auditory attention decoding an endpoint goal is the improvement of methods like beamforming to make them more accurate, robust, and adaptable.

1.4 Methods for Measuring Neural Data

In practical settings the neural response can be very complex due to aspects of the auditory scene and how the listener is reacts to it including extraneous noise, constant attention switching, reverberation, characteristically close input audio streams, etc. This has lead experiments to focus on simplified auditory scenes that consist on only two individual, spatially separate speakers introduced to a stationary listener or through headphones to the listener. These simplified scenes make it easier to attribute the ERP activity measured in the neural recordings to characteristics of speech. In approaching a blind-source separation task, such a basic auditory scene where the the multiple audio sources are separated makes it a much more manageable task. Recent experiments have found success in applying source separation to auditory scenes prior to comparing with the reconstructed audio (Han et. al, 2019). Furthermore, methods have been proposed for successful AAD tasks using the noisy reference audio that such source separation methods would normally produce (Aroudi et. al, 2016). In the development of this task each neural recording technique has shown its capability in decoding attended envelops from event-related potentials.

1.4.1 Electrocortiography (ECoG)

ECoG is a method of recording electrical activity in the cerebral cortex using electrodes placed directly on the exposed surface of the brain. Due to the invasive nature the technique can better capture the spectrotemporal structure of the neurological process of the brain. Using ECoG recordings to reconstruct audio envelopes from the broadband gamma range (70-170Hz) has been shown to be effective in determining attention in a two-speaker environment (Dijkstra, 2015). In fact this study also showed that significant accuracy can be achieved with as little as one electrode in use. However, for practical use the invasive aspect of ECoG recordings makes them entirely impractical for uses in commercial devices. The use of the ECoG recordings is then only really limited to those suffering from severe neuro-degenerative diseases that prevent the use of conventional assistant devices.

1.4.2 Magnetoencephalography (MEG)

MEG is a neuroimaging technique where the brain activity is recorded by monitoring the magnetic fields produced by the electric currents in the brain. Early studies using MEG were able to reveal a correlation between theta band (4-8 Hz) modulation at 200ms post-stimulus and intelligibility of speech signifying tracking how listeners hear is a possibility (Luo and Poeppel, 2007). In recent years MEG has been shown to be a reliable measurement technique in more advanced AAD tasks such as switch attention settings (Akram 2016, de Cheveginé, 2018). In the real-time applications needed for the accurate performance of hearing aids and hearing implants MEG recordings have a few shortcomings. First, MEG recordings cannot be taken without specialized, noisy equipment making it very impractical to integrate into such small devices. Second, MEG recordings provide very good spatial-resolution in separating cortical sources (Silva, 2013), yet lack in its ability to provide the same temporal resolution of EEG without the formally mentioned equipment. Recent developments have reduced the size of the equipment for obtaining MEG recordings, but it is far from portable.

1.4.3 Electroencephalography (EEG)

EEG is another noninvasive technique where the brain's electrical activity is measured with exterior nodes. These nodes capture the brain's spontaneous electrical activity, which produces neural oscillations. Due to the ease of neural data collection it is very easy to collect large data sets for training. In contrast to MEG recordings EEG has the temporal resolution to record the neural response to continuous speech (Lalor et. al 2009), yet it lacks in spatial resolution requiring many nodes to get the same location specific readings as MEG. It has also been shown that EEG is more sensitive to attention-related neural activity components (Kahkonen et. al, 2001). Further research has looked into the use of minimal nodes for obtaining the neural readings while still getting significant decoding accuracy (Fuglsang et al., 2017). Also, unlike the other methods, progress has been made in developing wearable unobtrusive devices that use EEG to solve AAD tasks (Looney et. al, 2010, 2011). The use of EEG is then more practical for AAD tasks and will be the method focused on for the continuation of this section.

1.5 Decoding Methods

For the reconstruction of the envelopes, represented as $s_a(t)$, a linear spatiotemporal decoder is has been shown to be effective for AAD tasks (Fuglsang et al., 2017). This decoder, defined as $g \in R^{N_l \times C}$ can be described through the following:

$$\overline{s}_a(t) = \sum_{n=0}^{N_l-1} \sum_{c=1}^{C} g(n,c) M(t+n,c) \tag{1.1}$$

$M(t,c)$ is used to represent the C-channel EEG recording at discrete time t and neural channel c. The time lag index is represented by n, which has time lags between 0 and N_{l-1} samples. The time lag is meant to represent the lag between the brain receiving the auditory stimulus and when the brain processes it. The most effective time lag has been shown to be around the 250ms range (O'Sullivan, Power, 2015). The reconstruction of the attended envelope at time t is the weighted sum of all C channels as well as future samples up to $t+n$. Applying a decoder to the EEG signals

is referred to as a backwards model throughout the literature. In contrast, a forward model is when a decoder is used on the stimulus envelope prior to calculating its similarity to the EEG recordings. This process is described through the following:

$$\overline{r}(t,n) = \sum_{l=1}^{L} h(n,l)S(t-l) \tag{1.2}$$

$h(n,l)$ is the forwards decoder while $S(t)$ represents the input speech envelope. Once the envelope is reconstructed a similarity measurement, such as correlation coefficients or mean-square error (MSE), between each identified input stream and the reconstruction is taken to determine accuracy. Difficulties do occur in the forward models as it is hard to replicate the noise in the neural recordings for a more percise comparison. From these descriptions it is clear that the forward model corresponds to bottom-up attention models while the backwards model is directly related to a top-down attention model.

1.5.1 Variations on Linear Decoding

For the decoder to optimally reconstruct the envelope much consideration must be given to the weights of the decoder. In a linear method a common implementation is a regularized, least-squares transform (LSQ) (O'Sullivan, 2015). The LSQ weights are referred to as the temporal response function (TRF). These TRFs characterize how the changes in EEG recordings correspond to a unit impulse in the input features. The TRFs have also been applied to decoding of speech spectrograms (Ding et al., 2012), phonemes (Di Libierto et al., 2015), and semantic features (Broderick et al., 2011). Regularization methods are necessary to constrain the model coefficients to prevent overfitting. In a comparative study between multiple forward and backwards models, each applying a different regularization method, it was found that backward models not only outperformed the forward models, but also benefited significantly more from specific regularization method (Wong et. al, 2018). Such a direct method for decoding either the stimulus from the whole neural data or the neural data from the stimuli is open to a lot of error, especially since the brain implements non-linear processing techniques. Other methods propose new ways to determine similarity to improve this aspect.

Canonical Component Analysis

Resulting similarity scores, specifically correlation, that linear methods use generally produce lower values ($r = .1 - .2$) due to the fact that EEG recordings account for a plethora of the brain's sensory processing, much less its overall activity. This makes it so the variance of the EEG data cannot be accounted for by the audio stimuli. Canonical Component Analysis (CCA) is a method that was introduced to reduce the redundant variation in the stimulus and EEG data by proposing the use of an optimal transform for both stimuli and EEG that leads to the highest correlation (de Cheveigné et al., 2018). In this study multiple CCA transforms were proposed and tested against a linear forward and backward model. It was found that the correlations reached their highest values when two finite impulse response (FIR) filters were used for the transforms of both the EEG data and the stimulus envelope. Computationally CCA was found to cost only a mild amount of memory to out-perform the linear tasks. Overall, CCA provides a significant improvement to the correlations of the EEG data and stimulus data with a computationally efficient technique. It has also been suggested that CCA would perform better when used with non-linear transforms of the data to take into account how the brain processes audio.

Statistics driven State-Space Models

Another proposed method is the integration of state-space-models that makes use of statistical measurements to determine the likelihood of attending to one speaker or another (Akram et al., 2016, Miran et al., 2018). A state space model in an AAD task consists of two components, one relating the neural measurement to a set of unobserved state variables, such as the forward model, and another describing the evolution of those unobserved states over time. The parameters for these components can be tuned for each new event when implemented in a probabilistic framework. Such a proposed framework includes a Maximum a-posterior (MAP) estimation, found through using an EM algorithm, being used to infer the state-space parameters while the attentional state is modeled by a statistical distribution (Akram et al., 2016). One key advantage is that these produced parameter estimations have confidence bounds that can be used in inference procedures like hypotheses testing. The same study found that the state-space model could provide highly

accurate results in the common two speaker setting with a reduced temporal resolution of under 10 seconds.

The state-space model has also open up the general method of attention decoding to interpret more complex processing of the brain, specifically how the brain is able to switch which audio stream it is paying attention to. In recent models a Bayesian representation of the state model has been implemented and is shown to operate in near real-time with high accuracy in both consistent and switched attention states (Miran et al, 2018). This study was able to tune the state space model to interpret the similarity scoring methods obtained through the decoding of both EEG and MEG recordings. Some noteworthy advantages include a temporal resolution of 1 sec for robust attention decoding performance and the need for only minimal offline training. The addition of the state-space model addresses many of the real-time applications issues that the decoders present.

1.5.2 Non-linear Models

In order to better model the auditory processing that takes place in the auditory pathway many non-linear models for training the weights of the decoder have been proposed. Machine learning methods make use of non-linear processing, which not only help account for the neural processing of the brain, but the compression and loss of fine structure information that also occurs. These can enhance the CCA method by taking the newly transformed neural data and envelope and creating a more robust mapping between them. Similarly, in the use of state-space model approaches the similarity scores used in the probabilistic framework would benefit from more robust decoding of the neural data. As of yet not much effort has been focused towards the combination of these methods in performing the AAD task. Rather the sole improvement over the decoding accuracy of the linear forward and backwards methods through the use of a single method has been the goal in recent years.

Deep Neural Network (DNN)

One such technique is the implementation of Deep Neural Networks. In automatic speech recognition tasks the integration of DNN and other non-linear models show improvement compared to prior state-of-the-art methods (Hinton et al., 2012). This gives evidence for the use of DNN topologies for replicating the auditory processing of the brain. DNNs are advantageous to a linear regression model due to the tunable parameters that give more precise reconstructions. DNNs have also been used to evaluate the relevance of inputs (Fuglsang et al., 2017). The study shows that DNNs can identify what neural activity is most relevant to the AAD task. This leads to crucial aspects such as reduction in computational complexity and the reduction in the amount of neural activity that needs to be measured. A DNNs nonlinearity also plays a role in reducing the error in the inverse mapping. When applied to AAD tasks a simple DNN containing minimal layers has been shown to perform consistently equal with the linear regression models (de Taillze et al., 2017, Ciccarelli et al., 2019). One of these studies actually shows improvements of the DNN when using broadband (1-32 Hz) EEG data rather than the common narrowband (2-8 Hz) usually associated with neural queues for audio processing. This shows how the nonlinear nature of the DNN topology also allows for more in depth processing of the neural data. Such results have motivated other nonlinear machine learning methods to be applied to AAD tasks to find methods that provide greater accuracy while keeping the replication of the auditory system.

Convolution neural network (CNN)

Convolution Neural Networks are widely used in image classification and are a preferred approach in recognition and detection tasks. Studies have shown positive results for seizure detection from EEG data through the use of a CNN (Mengni et al., 2018). In recent iterations of implementing CNNs for identifying the attended speaker have done direct classification rather than creating a reconstruction and using a similarity measurement (Cicarelli et al., 2019, Vandecappelle et al., 2020). This reasoning was shown to provide better accuracy than the simple DNN. This difference increased as test window lengths increased, which means the accuracy is higher when the network has access to more contextual information. As with other neural network methods the optimal

goal for implementing these models is getting the highest accuracy in the shortest time windows. There are also studies that focus on the detection of the locus of the speaker using a CNN model (Vandecappelle et al., 2020). It was found to have a median accuracy of 81% with one-second decision windows. This gives an alternative approach to improving real-time adaptability of hearing aids through faster and more accurate beamforming once the direction of attention is determined. Until the accuracy in smaller decision windows is increased for stimulus reconstruction methods it is hard to find practical implementations.

Recurrent Neural Network (RNN)

For many tasks like language modeling and speech recognition Recursive Neural Networks have been increasingly popular due to how their topology has a feedback structure similar to how the brain is proposed to function. A variation of interest is a Long Short-Term Memory (LSTM) network. LSTMs are set up so that it has the ability to add new information to its memory, but also forget past information solving a standard RNN's long-term dependency issues. Currently little has been done to integrate this model into the decoding of neural data for AAD tasks. However, recently a model was proposed to relate small temporal segment EEG data to corresponding stimulus envelopes through transforming both to a common embedded space through the duel use of an LSTM in the speech path and a CNN in the EEG path (Monesi et al., 2020). In classifying whether a given stimuli corresponds to a given EEG set the study shows significantly higher accuracy than other methods. This gives reason to consider this approach as a comprehensive representation of EEG and speech for future AAD implementations. There also has not been as much investigation into the computational complexity of an RNN model when compared to previously mentioned nonlinear methods.

1.6 Current Applications

Currently millions of people are suffering from some degree of hearing loss (Wilson et al., 2017). Hearing loss more commonly affects older adults and is associated with other signs of advanced age

as cognitive decline and initial onset dementia (Dawes et al., 2015). Furthermore, Veterans are also disproportionately effected by hearing loss (USVA, 2017). It has been shown that when people are not able to actively participate with those around them it can lead to social isolation (Mick et al., 2014) and an increase in depression (Mener et al., 2013). It then becomes essential in maintaining quality of life to make use of a hearing assistance device. Critical flaws in the functionality and design of commercial devices can actively make people choose a lesser quality of life rather than deal with ineffective solutions. The inability for these devices to enhance audio streams of interest can lead to such negative experiences that many decide not to use them. Also, a bulky or invasive design of the devices can have a negative impact on the user both practically and aesthetically. This gives even more reason to try and find an effective, portable method of preforming AAD in a real-time complex environment. As of now the collection of EEG measurements is by far the easiest measuring technique to record ERPs for the purpose of implementation into smaller, portable devices. In fact, much progress has already been made in developing portable EEG systems (Sawan, et al., 2013). A current major roadblock in the commercialization of this is maintaining the accuracy while operating in real-time and constantly introduced to new situation with little training data.

In the majority of research there is little investigation into how the environment impacts the performance of AAD methods. So far the performance of AAD methods in a duel-speaker setting is respectable, yet practically it is very rare to be in an environment free from reverberation or extraneous noise. A recent study has made progress in this by evaluating the accuracy of a two speaker AAD task using EEG recordings when such reverberation and extraneous noise are introduced (Aroudi et al., 2017, 2019). The study revealed that when the environmental conditions for the training and testing sets are the same equivalent decoding accuracy to the anechoic condition could be achieved. Furthermore, decoders trained on multiple conditions are found to have significant increases in accuracy than just training on a single condition. It is then promising that the methods applied in anechoic environments can still be easily translated to more complicated environments.

1.6.1 Hearing Aids

The most prominent application for AAD is the direct improvement of wearable assistant devices for hearing loss. Common hearing aids generally solely work on applying the technique of beamforming to selectively enhance part of the input audio without concern for what the user actually wants to attend to. Therefore the user's actual intention is not taken into consideration and the current methods rely on essentially a guess for what people want to pay attention to. In terms of these devices the direction of enhancement for the beamforming is either set to where the wearer is looking or the direction of whatever input stimuli is the loudest. As stated previously this can cause a lot of extraneous noise to be amplified. Many patients of hearing loss then actively choose to not use these devices as it leads to even more frustration. There are also issues with the size of the devices. It has become common to see advertisements for hearing aids that place as many components directly in the ear canal to effectively hide the device. The proposed real-time data collection techniques face a challenge when integrating the processing tools they need while not reverting to the bulky design many users to not prefer.

1.6.2 Cochlear Implants

For hearing loss that is caused by damage to the cochlear mechanisms of the inner ear a cochlear implant is a common option. The implant consists of two parts, an outer microphone and speech processor and a surgical implant. When the microphone receives audio signals they are processed by a digital signal processing (DSP) unit before being sent to the internal implant. This implant converts the received speech signals to electrical impulses that are used to stimulate regions of the auditory nerve. The stimulation is done by wires threaded into the cochlea, which have electrodes to have direct access to the auditory nerve. The goal of these implants is to ultimately recreate the signals sent to the auditory nerve that the outer and inner ear mechanisms would create when in the same auditory scene.

It has been documented that the performance of listeners with cochlear implants in performing sentence recognition vastly improves with the implant (Zeng et al., 2008). At the same time these devices leave much to be desired. In a single speaker setting a much higher signal-to-noise ratio

(SNR) is required to achieve 50% successful sentence recognition for implant users than normal listeners. This difference becomes even greater in multi speaker setting (Zeng et al, 2008). Furthermore listeners with implants are significantly impaired in simple melody and tone recognition tasks. This alludes to other evidence that indicates that cochlear implant users suffer in speaker identification tasks (Vongphoe et al., 2005). In practical applications these implants may give back key functionality, but leave the users without crucial auditory resolution.

1.6.3 Brain-Computer Interfaces (BCI)

Both hearing aids and cochlear implants focus on delivering an interpretation of the incoming auditory scene to the auditory nerve and ideally the information received is enough for the brain to accurately decode the signals. By doing this there is no feedback from the listener, which would undoubtedly improve what information the methods should deliver to the auditory nerve. For those that suffer from such degenerative diseases that the use of the aforementioned devices is impossible brain-computer interfaces (BCIs) are used in communication tasks. BCIs are a direct example of the use of neural recordings to perform the desired task of the user. The communication task applications of BCIs involve decoding the sensory activity the user wants to perform. This has direct relation to every currently proposed method for AAD tasks since they all rely on the accurate decoding of neural data. The integration of BCIs into other hearing assistance devices is a key component in actually replicating a user's desired performance. Many researchers are focused on creating a functional feedback method from a user's brain activity to supplement the limited functionality in current commercial offerings.

1.6.4 Commercial Home Devices

So far applications have been presented that have focused on the fixing of lost sensory ability through replicating a natural neurological function. Beyond just helping humans AAD models can be applied to the enhancement of widely used technology. Many homes today have several devices that have voice activation capability. In such devices like the Google Home or the Amazon Alexa there is the ability to process the input audio and have an appropriate reaction. As stated, being

able to realistically process audio requires being able to function in complex auditory scenes. In these settings it becomes important for these devices to figure out what is important to listen to. Many methods for attention decoding can be applied to this task. Commonly the devices are trained on a single users speech and a comparative technique is used to tune what is important. Though this is a more basic implementation of AAD techniques there is clear applicability to mimicking human attention in these machines to listen for important information directed at it. Though these devices will likely always lack the intrinsic learning ability of a human it can still greatly benefit their performance.

1.7 Proposed Work

Current research has lacked in many aspects for developing real-time, adaptive models. In performing any task, especially those expected to have a near instantaneous reaction, it is key to prioritize computationally simple models. Another aspect is the avoidance of degradation in accuracy when presented with significantly less training data. Both of these have been a priority in the approach of many of the methods I have mentioned thus far. Despite the many methods proposed there has been little work done on testing the performance statistics of each against one another. This is even apparent in papers that explore the same decoding methods and implement the same performance metric. A main cause is that many factors in the experimental procedure vary. This includes the number of speech streams presented to the listener, what angles these speakers are placed at in relation to the listener, and the gender of the presented speakers. Differences in data collection can also lead to an inability to justifiably compare methods. EEG electrode placement also tend to vary across various publications as there are multiple standards that can be used in EEG data collection. Variations are also present in the processing parameters, which includes the bandwidth of the EEG or speech-envelope used and the temporal context used for reconstruction.

To better address the issues of variability brought up among multiple methods this thesis presents a simple DNN model presented by (Ciccarelli et al.,2019) and applies a data set from (Fuglsang et al. 2017). The data set being integrated into the DNN consists of trials where a listener was presented with two opposite gender speakers each telling a different story and asked to pay consistent attention

to a single speaker. The original data set also had listeners presented with the same dual-speaker attention task. Also, the collection of the EEG data and the processing done on both data sets is very similar. The two do vary in the method with which they were played to the listener. (Ciccarelli et al.,2019) presented the two co-located speakers from a front facing loudspeaker. In (Fuglsang et al. 2017) the listeners are presented with the spatially separated speakers over earphones. In terms of the size of the new data set being integrated into the DNN had much shorter trials leading to a drastic decrease in the amount of data that can be used in the training and testing steps. Due to the reduction in the size of the data set the testing metrics also needed to change so that smaller reconstruction segments were used for determining accuracy. This was also practical in allowing for enough data to be given to compare the performance of the network on individual listeners. In both methods the predictions were done using a correlation metric, again making it easier to compare the results from each paper to the performance of the integrated data set into the DNN. In the next section the methods used to perform the AAD task are described in further detail.

By using a smaller data set the consistency of the simple DNN method when presented with less training data is explored. As presented the DNN network has been shown to produce results that reached near equivalent classification accuracy as a linear method used on the same data set. With less training data if the resulting prediction accuracy remained consistent on both an overall and individual metric then the network can be said to be robust to a significant decrease in available training data. This also gives the linear method explored in (Fuglsang et al., 2017) a comparison to a non-linear method. On an individual listener basis specifically, we would be looking to see if the training of the DNN is more preferential to certain listeners than to others. As with (Ciccarelli et al.,2019) this shows how the network can be applicable to multiple different users, especially since the newly implemented data set is comprised of more subjects.

The performance of this test shows how the comparison of AAD methods can start to take place. As more investigations are made into the comparison of multiple techniques many aspects of the models can be seen. One such example is that it can highlight the different strengths and weaknesses each have in performing an AAD task. Also, If methods are shown to be robust when presented with changing data sets then there can be a larger focus in determining which models should be focused on. As the accuracy of the models increases applicability becomes a much more prominent

issue. There is little use in reaching the best replication given it can not practically function both in a portable device and in a real-time, complex environment. A result this exploration is a start to the conversation on the feasibility of current state-of-the-art methods as they are further developed and improved upon.

Chapter 2

Methods

2.1 Data Set

The data set for this experiment was taken from (Fuglsang et al. 2017). The stimuli were made from the individual anechoic chamber recordings of one male and one female storyteller. They were asked to read stories and the resulting recordings were split into consecutive 50-second long segments. Two of the segmented speech streams, each from a different storyteller of a different gender, were presented to the listener at the same time from different spatial directions. Through generated binaural impulse responses of an anechoic room each speaker was placed at ± 60 degrees from center and at 0 degrees from the azimuth. There were also filters created for moderate and high reverb room conditions generated from simulated rooms. The placements of the speakers in these conditions are the same as in the anechoic room. Only in the linear method were the performances on the reverb conditions explored. They were presented to the listener at the same sound pressure level (SPL) of 65 dB. Data collected from a listener presented with a full 50-second segment corresponded to one trial and there were 20 trials presented to each listener. For all 18 listeners present in the data set every trial had the stimuli presented to the listener over insert earphones within a listening booth.

Prior to the presentation of the competing audio streams in a trial the listeners were told which

audio stream to pay attention to. In order to decrease variability and interference of other sensory attention cues the listeners were also instructed to minimize movement and fix their eye gaze for the duration of the trial. For each trial the position of the target speaker relative to the speaker and the gender of the target speaker were randomized. Furthermore, the order of which stories were presented to each listener were also randomized. This was done to minimize the likelihood that a speaker's identity, story pairings, or speaker's position influence the neural response to either the attended or unattended speech. In order to assess if the listener successfully attended the desired audio multiple choice questions were used as indicators. If the desired audio stream was attended, then the listener should be able to correctly answer a majority of the comprehension questions. The EEG recordings were recorded using 66 electrodes mounted on a head cap with a 10/20 layout. Each electrode was sampled at 512 Hz.

2.2 Data Pre-processing

The individual signals from each electrode were band-passed filtered with a zero-phase forward filter. The data for each channel was then downsampled to 128 Hz and re-referenced to the average response of the 2 extra electrodes (65/66) on the mastoid. These two extra electrodes on the mastoid act as reference points so that the re-referencing is not based on any of the main 64 EEG channels. Additional processing was done to remove channels and components with excessive noise. Any removed channels were re-interpolated using spline interpolation. Finally, the EEG recordings were low-pass filtered at 8 Hz using a zero-phase second order Butterworth filter. This was done to capture the most prominent frequency ranges for neural tracking of speech envelopes, the theta (4-8 Hz) and the delta (1-4 Hz) bands. Both the EEG data and the extracted speech envelopes were downsampled to 64 Hz. This helped in keeping the computation of the network reasonable while maintaining an effective temporal resolution.

2.3 Reconstruction Methods

A backward attention model was used to determine the attention of the listeners. In Figure 2.1 the process for performing an AAD task using a backwards model is depicted. In a backwards model a temporal response function (TRF) was used to reconstruct the clean attended stimulus envelope from the recorded EEG data. Then the reconstruction of the clean attended envelope was individually compared with the two clean input audio streams. The correlations were calculated using a Pearson's correlation coefficient. A correct classification of determining attention was if the reconstructed stimuli had a higher correlation coefficient with the attended stream envelope than the unattended stream envelope.

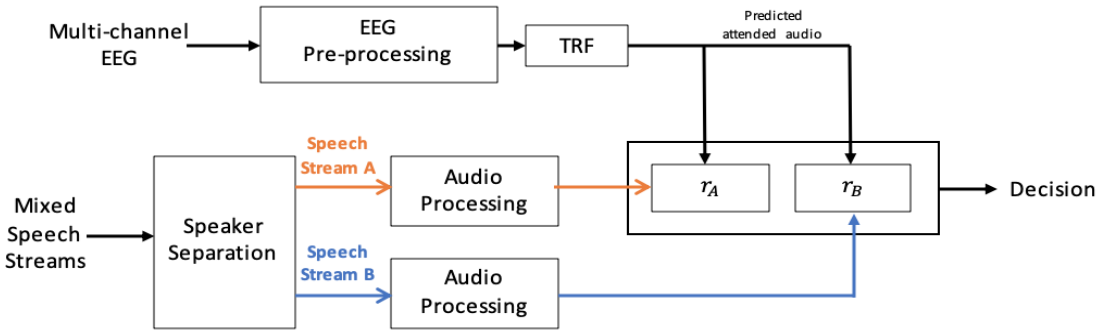


Figure 2.1: Backwards model system architecture for auditory attention decoding in a dual speaker environment

2.3.1 Linear Method

The linear method used was taken from (Fuglsang et al., 2017). To reconstruct the attended audio stream envelope from the EEG data a finite impulse response (FIR) filter is trained to perform linear mapping between the two. This spatio-temporal decoder was described in equation 1.1. The filter coefficients were estimated using ridge regression, which was trained on the data present. Two backwards decoders were trained, one for decoding the attended stream and the other for decoding the unattended stream. Both used a leave-one-out paradigm where the same trial from all listeners was left out of training and kept for the testing set. Roughly there were 15 minutes of testing data

in total and 4.75 hours of data to train the model when the full data set was in use.

The accuracy of the reconstructions were determined based on the Pearson's correlation coefficient as follows:

$$r_p = \frac{cov(\hat{e}_a, e_a)}{std(\hat{e}_a) \cdot std(e_a) + \epsilon} \tag{2.1}$$

Here $\hat{e}_a$ refers to the reconstructed attended audio envelope and e_a represents the corresponding section of the actual attended audio envelope. The correlation was determined as the covariance between the two segments over the product of each segment's standard deviation. ϵ refers to some small value (10^{-30}) that is used to prevent any divisions by zero in rare cases. The decoding was determined successful for a segment of the reconstructed envelope if a higher correlation was achieved with the attended audio envelope rather than the unattended audio envelope.

In learning the reconstruction filter the weights were fitted using multiple post-stimuli time lags of EEG data were used. This ranged from 0 ms to 500 ms and were used to determine the contributions of individual electrodes. Using this metric the least relevant electrodes across all time lags were identified and subsequently removed from training. This technique was done to reduce the noise that is prevalent in the EEG data and reduce the complexity of the training since less input data is needed to reach the same decoding accuracy. The reduction in EEG channels used was only a metric for the replication of the linear method. Similar channel reduction was not applied to the DNN.

2.3.2 Deep Neural Network

To better replicate the non-linear acoustic signal processing of the auditory pathway a Deep Neural Network (DNN) was used. The network architecture was modeled after the one proposed in (Ciccarelli et al., 2019) where the network consisted of one hidden layer and two nodes. Figure 2.2 shows a detailed visualization of the hidden layer architecture. This part of the network is specifically inspired by the design presented by (Taillez et al., 2017). To reconstruct a single time point the temporal context used is the EEG data from that time point and the 15 consecutive time points

in the future. This corresponds to 250 ms of post stimulus lag, which is done to account for the delay in the processing of the brain. Other features of the audio are then still present in the neural data at later time points. This operation is performed over all the time points in a predetermined prediction window. In experiments this was chosen to be 16 time points. Once the 16 time points were reconstructed a correlation loss function was applied between the full reconstructed prediction window and the corresponding section from the attended audio. The correlation loss function was used to maintain consistency with how the testing accuracy will be evaluated using a correlation metric as follows:

$$C_{corr}(\hat{e}_a, e_a) = 1 - \frac{cov(\hat{e}_a, e_a)}{std(\hat{e}_a) \cdot std(e_a) + \epsilon} \tag{2.2}$$

Here A resulting cost of zero implies that there is perfect correlation between the predicted and attended audio stream. Alternately, a cost of one signifies no correlation. The weights of the network are then updated to minimize the cost function. In minimizing the cost function the correlation should be gotten as close to +1 as possible. Correlation loss above one refers to negative correlations which the minimization causes the correlation to move away from. The loss functions to ensure that the reconstructed audio has a strong positive correlation with the attended audio. The average loss over the entire training was calculated at each epoch and used as a metric for the success of the training based on its resulting slope in later epochs.

Further modifications to this model were proposed by (Ciccarelli et al.,2019) This includes batch normalization before each layer, a non-linear activation function, dropout from the hidden layer, and a H hyperbolic tangent (Htan) activation function in the output layer to bound the output values. Batch normalization is used to limit the covariance shift in data. For EEG data having a higher dimensionality than the desired output this can greatly speed up the training. A non-linear activation function was then used to model the signal processing of the auditory pathway. A dropout chance of .25 was implemented to account for any over-fitting that may occur. Finally the Htan function acts as the output layers activation function to keep the output values between -1 and 1 which corresponds to the bounded input audio streams. The network and its organization are shown in full detail in figure 2.3. A critical aspect of the model is that significant compression

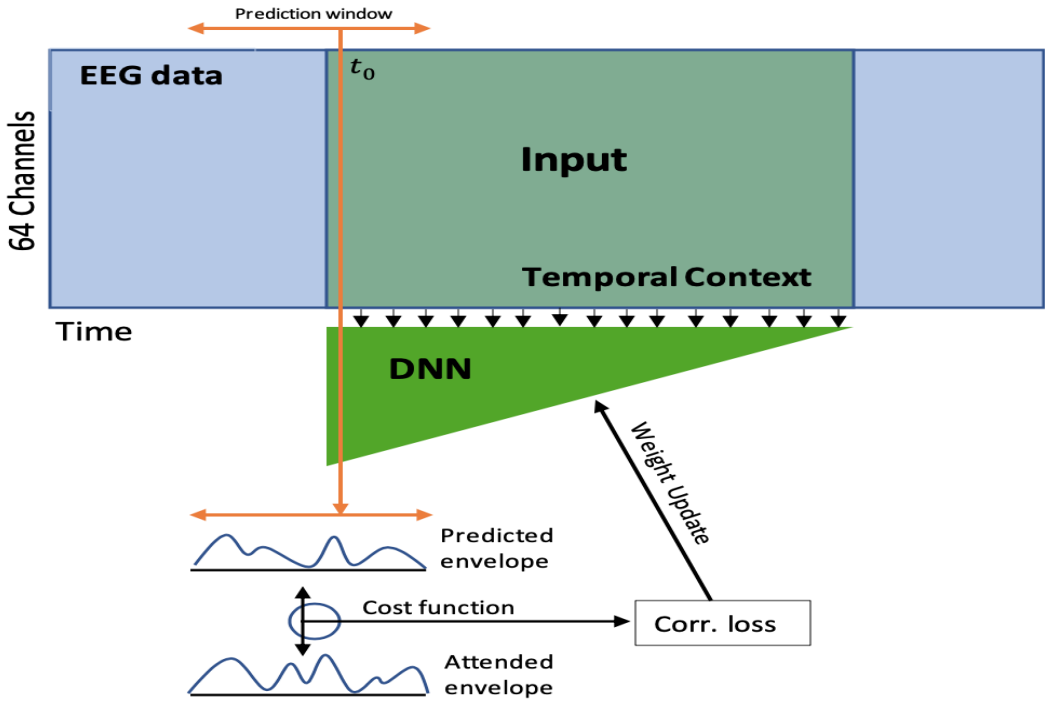


Figure 2.2: Structure of the Deep Neural Network layer as inspired by (Taillez et al.,2020)

of the EEG data is enforced to match that of the input audio. An Adam optimizer was used with a learning rate of 10^{-4}. Previous work used a learning rate of 10^{-3}, but the proposed reduction in training data required a smaller learning rate.

The training of the model consisted of a leave-one-out paradigm to maintain the consistency with the linear model. Such consistency allowed for more reliable comparison between the two methods results. The optimization was done using an Adam optimizer with a mini batch-size of 1024. In these trials for the testing set the listeners were all listening to the same condition and attending to the same stimuli. This allows the network to show its effectiveness for training on individual users in a new, consistent environment. The training was done over a run of 250 epochs where the total accuracy and loss were calculated after each epoch. The weights from each epoch were saved and used for the initialization of the subsequent epoch. The amount of epochs was chosen based off of past work and multiple trials where the decreasing curve of the total average correlation loss among the reconstruction started to consistently flatten.

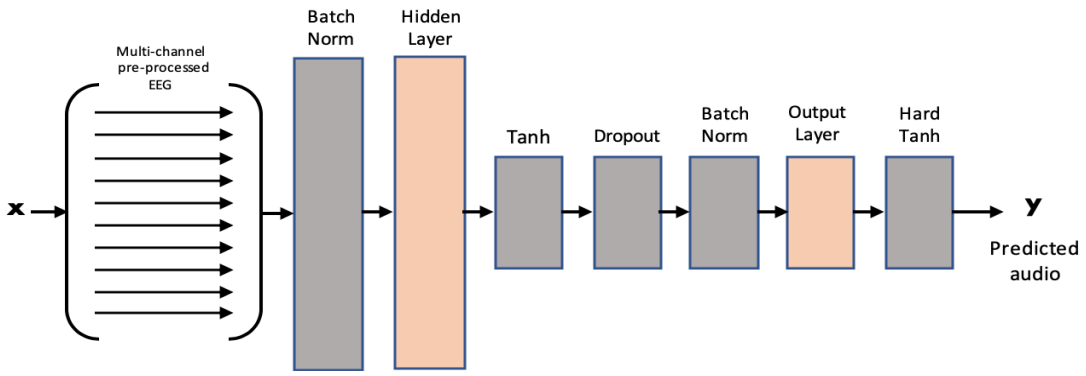


Figure 2.3: Full Deep Neural Network architecture as presented by (Ciccarelli et al., 2019)

2.3.3 Testing metrics

For the testing procedure the entire 50-second trial of audio was reconstructed given the corresponding EEG data. The reconstruction was then broken down into ten 5-second segments. As the DNN depended on a temporal context of 16 time segments for each time point reconstruction the last 16 time segments of the trial can not be reconstructed from the same amount of information. Therefore, the tenth segment is left out of testing to maintain consistency among all segments. With the remaining nine segments the correlations between each of them and the corresponding 5-second segment of the attended and unattended audio streams were calculated. The accuracy of both the total average among all listeners and the average of each individual listener was examined.

The accuracy of the total model over each epoch acted as an effective means to tell if the network was properly learning. A positive increase in the total accuracy accompanied by the opposite trend in the average loss of the network showed signs of there being learning across all subjects. The breakdown into each individual accuracy showed if the model was consistently under-performing for a specific listener across trials. If this was the case then the data corresponding to the listener was considered for removal from the training and testing set. Just like previous findings it was thought that the model would not perform equally across all trials and that the worst listeners could possibly influencing the training.

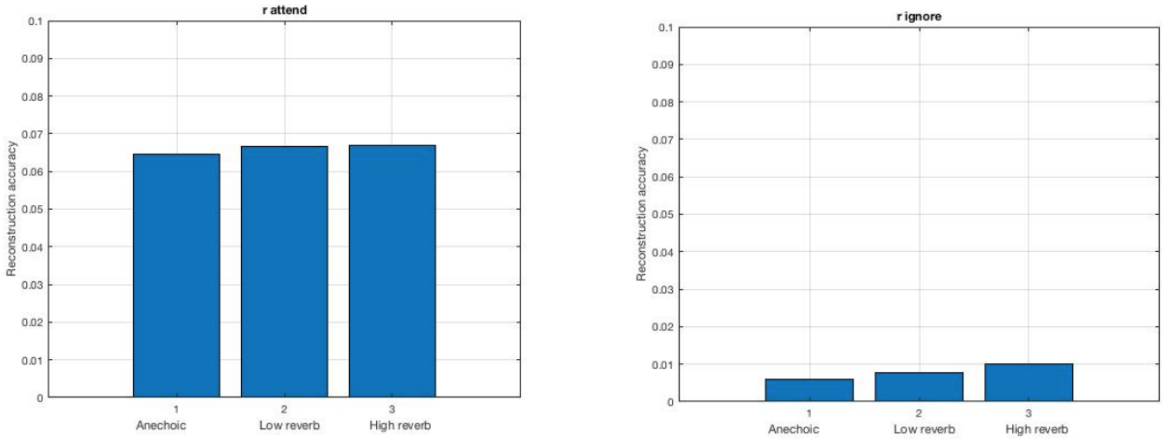
Chapter 3

Results

3.1 Linear Method

The replication of the Linear Method presented in (Fuglsang et al., 2017) helped frame the performance of the data set when applied to the DNN. The results are based on the overall data set rather than by the individual listener and are separated by the three environmental conditions that the speakers were presented in. Figure 3.1 shows the resulting average correlation coefficient of the reconstructed envelope with the envelopes of the attended and unattended envelopes. Through all acoustic conditions the reconstructed envelope on average produced a higher correlation with the attended audio envelope than the unattended audio envelope. This pattern is consistent with the results found in (Fuglsang et al., 2017), yet the behavior of the conditions differ slightly. Specifically the low and high reverb condition trials showed higher correlation averages with the attended audio envelope than the anechoic condition trials did. This however was not a significant difference between the three. For the correlations with the unattended audio envelope the averages were also consistent with the previous findings. The difference between the average correlations indicates that the replication was accurately training the reconstruction filters to model the attended audio.

When these results are broken down by individual trials the behavior continues to hold and no clear visible outliers are skewing the data. Figure 3.2 presents the comparison between the calculated



(a) Avg. Correlation with attended audio envelope (b) Avg. correlation with unattended audio envelope

Figure 3.1: Avg. correlation with input audio stream envelope

average correlation between both input audio stream envelopes for each trial. Nearly every decoder, no matter the condition or trial, had a higher average correlation with the attended audio envelope. The anechoic and low reverberation trials had less variability than the high reverberation trials which may account for the difference in overall average correlation values. Nonetheless the pattern for the individual trials is again consistent with the findings of (Fuglsang et al., 2017).

For the resulting decoding accuracy the results were able to be successfully reproduced. Figure 3.3 shows these average decoding accuracy for each acoustic condition. The performance for each condition remains consistent with previous findings despite any differences in the calculated correlations. The anechoic condition reached a decoding accuracy of around 85% while the low and high reverberate conditions reached around 82% and 76% respectively. The higher accuracy of the anechoic condition shows consistency with the average correlation difference between the unattended and attended audio envelope. The lower accuracy achieved in the reverb conditions are another aspect that is consistent with previous findings. Even more consistent with the previous findings was the exploration of the average filter weights at multiple time lags. Peaks in the attended decoding filter weights for all three conditions were found at around 350 ms. This corresponds to the previously presented correlation coefficients. No such consistent peak was found for the unattended filters. Though the pattern is consistent previous finding found this peak at 188 ms and the inconsistency has not yet been accounted for.

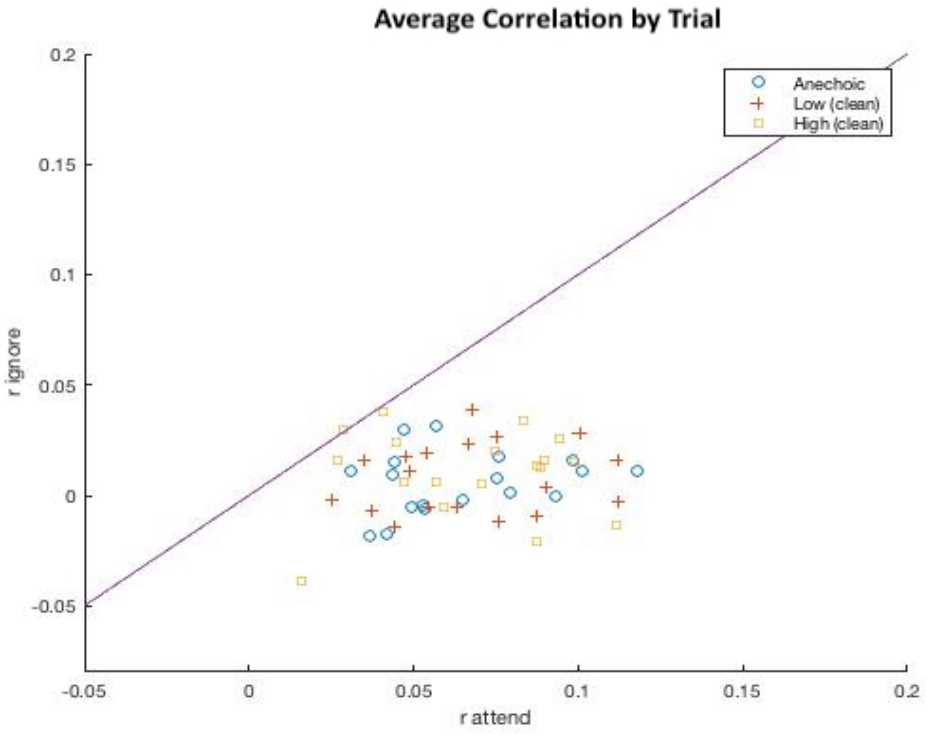


Figure 3.2: Comparison of average correlation by trial

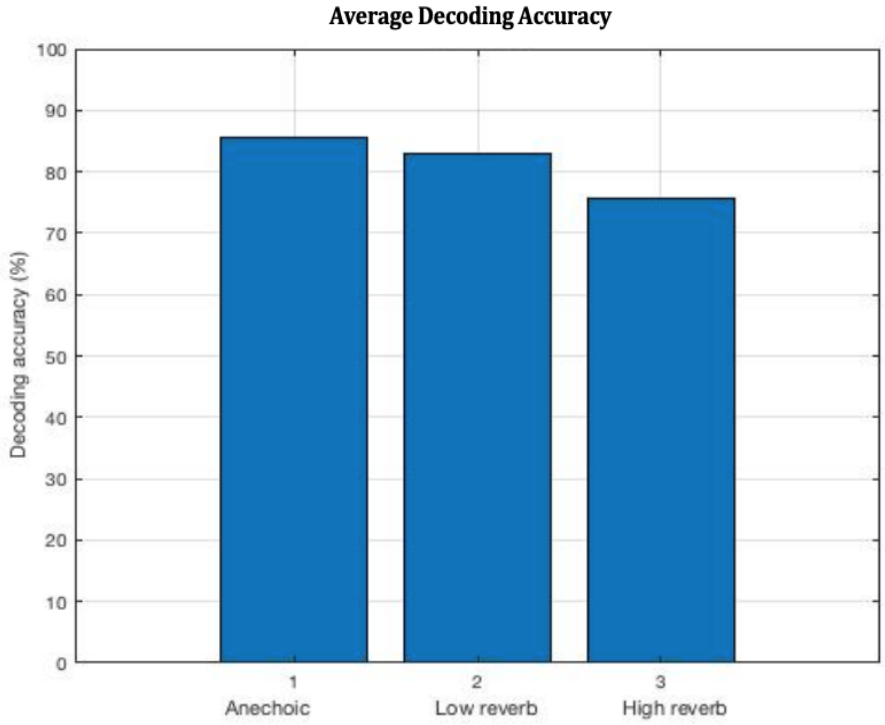


Figure 3.3: Comparison of the overall average decoding accuracy across room conditions

In implementing the linear method in (Fuglsang et al., 2017) many of the same patterns across all acoustic conditions were found. This pertains to correlation coefficients, overall accuracy, and a clear peak in decoding filter coefficients. These results signify that the results presented by (Fuglsang et al., 2017) are proper comparisons to the findings of any alternative method applied to the data, included the proposed non-linear method.

3.2 DNN Reconstruction of Attended Envelope

For each 5 second segment of testing data the reconstructed audio was made up of individually constructed segments of 16 time segments. In figure 3.4 the full reconstructed envelope for a testing time segment is shown next to the corresponding attended audio. The reconstructed segment is the full result of training on the full data set. It can be seen that the reconstructed segment has a sinusoidal nature to it, which is preferable for creating an approximation for speech which is known to have sinusoidal characteristics. The weight updates of the model depended on increasing the correlation of these two so it can be seen that maintaining the clear sinusoidal pattern was a factor in doing this.

3.3 Full Data Set

The application of the full data set allowed the performance of the proposed nonlinear neural network to be tested. For nearly all trials the average loss across all listeners generally presented in a consistent decreasing trend. Figure 3.5 shows many of the resulting loss graphs corresponding to which trial across all listeners was used as the testing set. These trials were chosen to showcase the variations in the performance of the network. The trial number for each average loss graph refers to the trial taken from each listener that was used to make up the testing data. Exponential functions were fit to each curve to better assess the overall trend of the loss function. The general success of the loss function was determined by its consistency and whether it still maintained a negative slope after 200 epochs. Certain trials were found to have inconsistent average loss trends or ones that ended in an increasing trend. These could be a result of improper training on the trial,

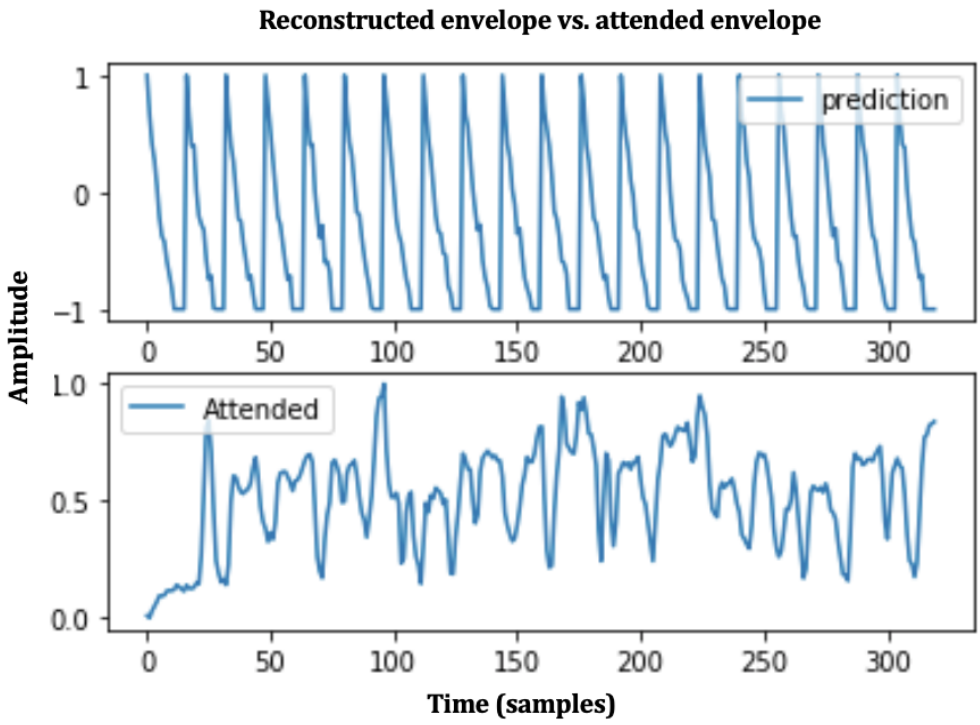


Figure 3.4: Sample of reconstructed envelope over a 5 sec. segment

over-training due to the small data set, or due to bad trials from the listeners.

To further investigate this the resulting accuracy plots were also looked at. Figure 3.6 presents the corresponding average accuracy charts over all subjects for the average loss charts in figure 3.5. Again a linear trend line was fitted to better assess the overall learning capability of the model. In this figure it is clear that the overall accuracy was maintaining around 50% no matter which trial was the testing set. Certain trials did show a slight increasing trend, but the variability kept the trend line under random chance. It was difficult to breakout the reason for this lack in performance as it was considered that the network may just be unable to train with much less data. Given that the model in (Ciccarelli et al.,2019) was trained on fewer subjects it was considered that the accuracy should be looked at on an individual basis.

There was a possibility that individual listeners could possibly be negatively influencing the training of the data. For each individual network tested on a different trial the individual performance of the decoder on each listener was examined. The criteria to be considered a bad performance was

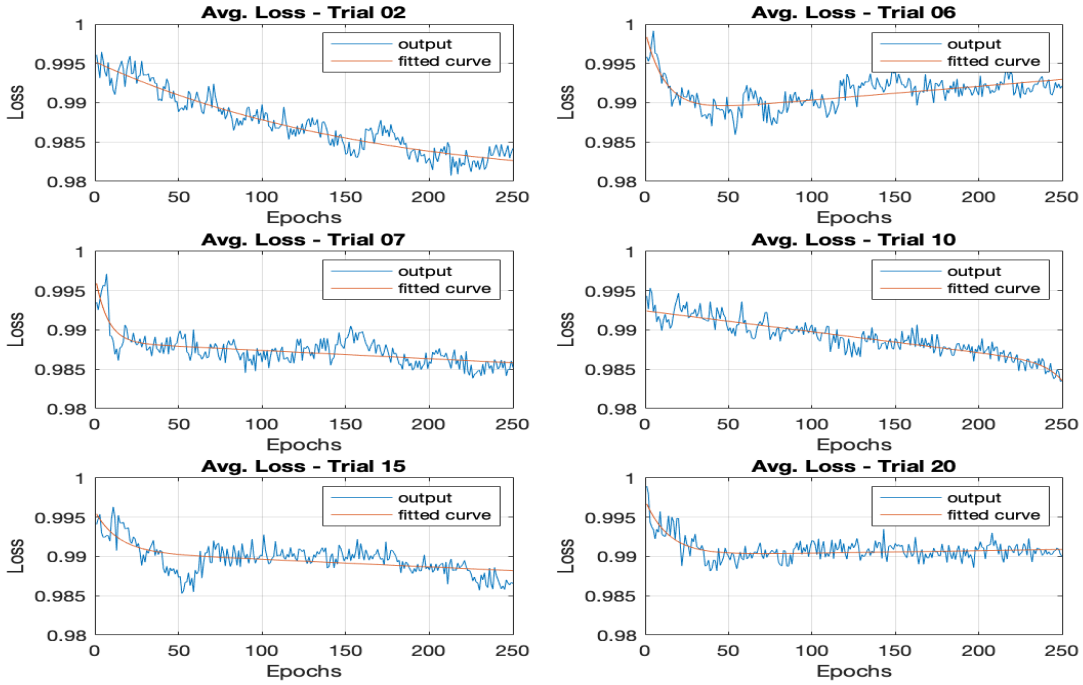


Figure 3.5: Average loss plots with all listeners across training epochs

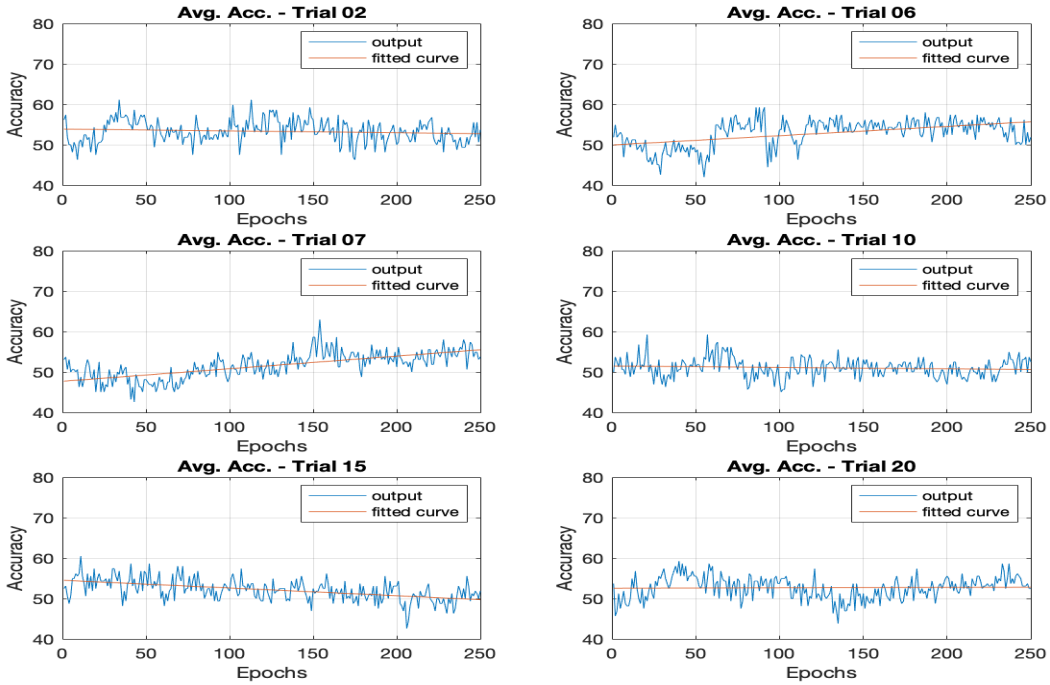


Figure 3.6: Average accuracy plots with all listeners across training epochs

the accuracy being below 45%. Over all 20 trained networks the six listeners that consistently performed the worse by this standard were then removed from training and testing data. By this set constraint the data of listeners 1,4,5,6,8, and 17 were removed from the data set prior to retraining leaving data from 12 remaining listeners.

3.4 Modified Data Set

With the update of the modified data the networks were retrained. This trade off of less data resulted in the training data being reduced to a total length of 3.17 hours and the testing set was reduced to around 10 minutes. This reduction further limited the data but clearly resulted in an overall increase in the testing accuracy. Figure 3.7 shows the resulting average loss functions of the reduced training data corresponding to the full training data loss functions presented in 3.5. Between the two trials 6,15,and 20 saw an improved behavior towards a negative trend line. Again, the trials that show a positive slope such as trial 10 were considered to be not effectively trained. The only trial that showed a variable pattern in the average loss was trial 7, but the overall slope of the fitted line was negative.

With the changes in the consistency to the average loss trends the resulting accuracy trends across the board also improved. The corresponding average accuracy plots to the loss functions in figure 3.7 are presented in figure 3.8. In the majority of trials the accuracy trends could be predicted from the average loss charts. This means that a consistently decreasing trend in a loss function pertains to a positive trend in the accuracy. It can be seen that less effective training occurred when trials 7 and 10 were used as testing data due to the negative trend found in each accuracy graph. This is in line with the less preferable performance in the average loss functions. Overall, training values were consistently performing around 60% by the end of the training, which shows that the removal of certain listeners improved results by about 10% regardless of the decrease in the training data. It should be noted that certain trials seemed to stagnate in terms of accuracy by around Epoch 100 causing a rather flat trend. One such example is shown in trial 15. Here the corresponding loss curve has a consistent negative trend, yet training accuracy overall flattened out. However the testing accuracy remained just over 60%, which is above the chance level performance.

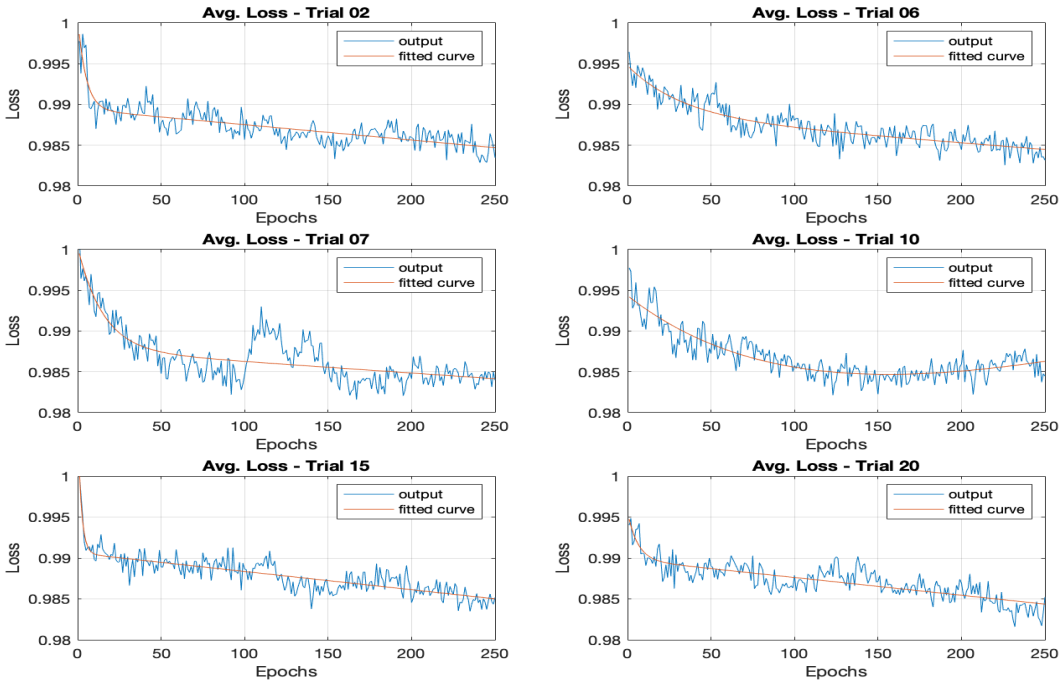


Figure 3.7: Average loss plots with reduced data

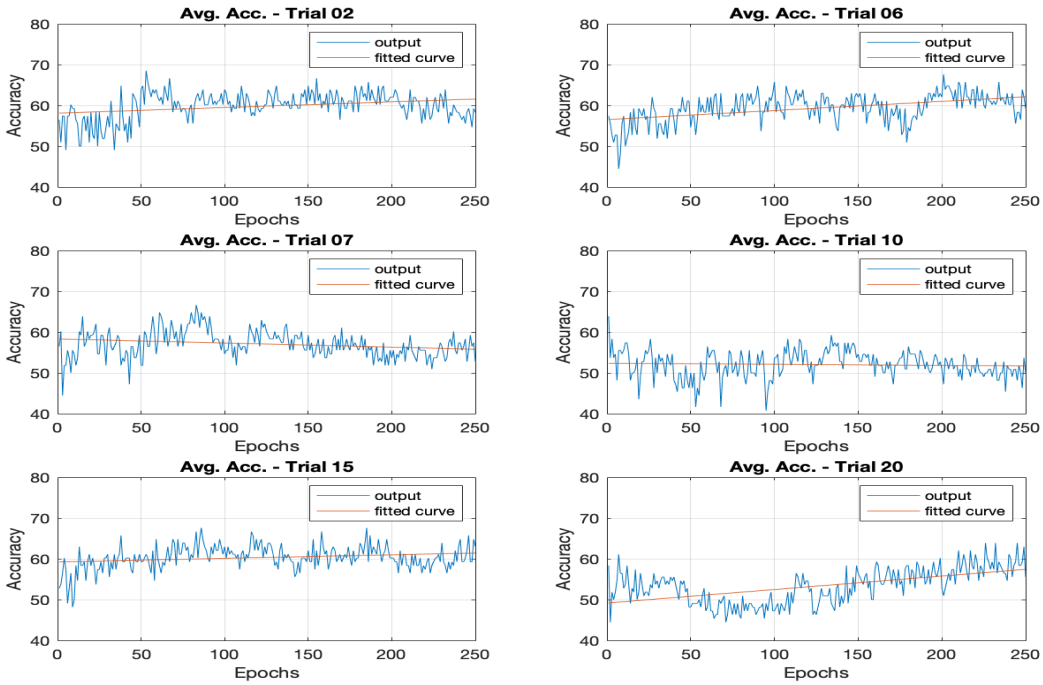


Figure 3.8: Average accuracy plots with reduced data

Previous work presented in (Cicarelli et al.,19) measured the decoding accuracy by listener. To better compare the results of the simple DNN with that of previous work the decoding accuracy for individual listeners are presented. Figure 3.9 shows the breakdown of decoding accuracy by listener for the six trials that corresponds to those shown in figures 3.7 and 3.8. The chance level prediction line is also included as a marker for performance. The behavior of the decoding accuracy in each trial again reflects the behavior of the loss curve. Specifically, for trial 10, which had an increasing loss trend, more listeners were performing under chance level. Trials that had a negative loss curve had more decoding performances at or above chance level. Previous findings also saw variability in the performance across listeners where

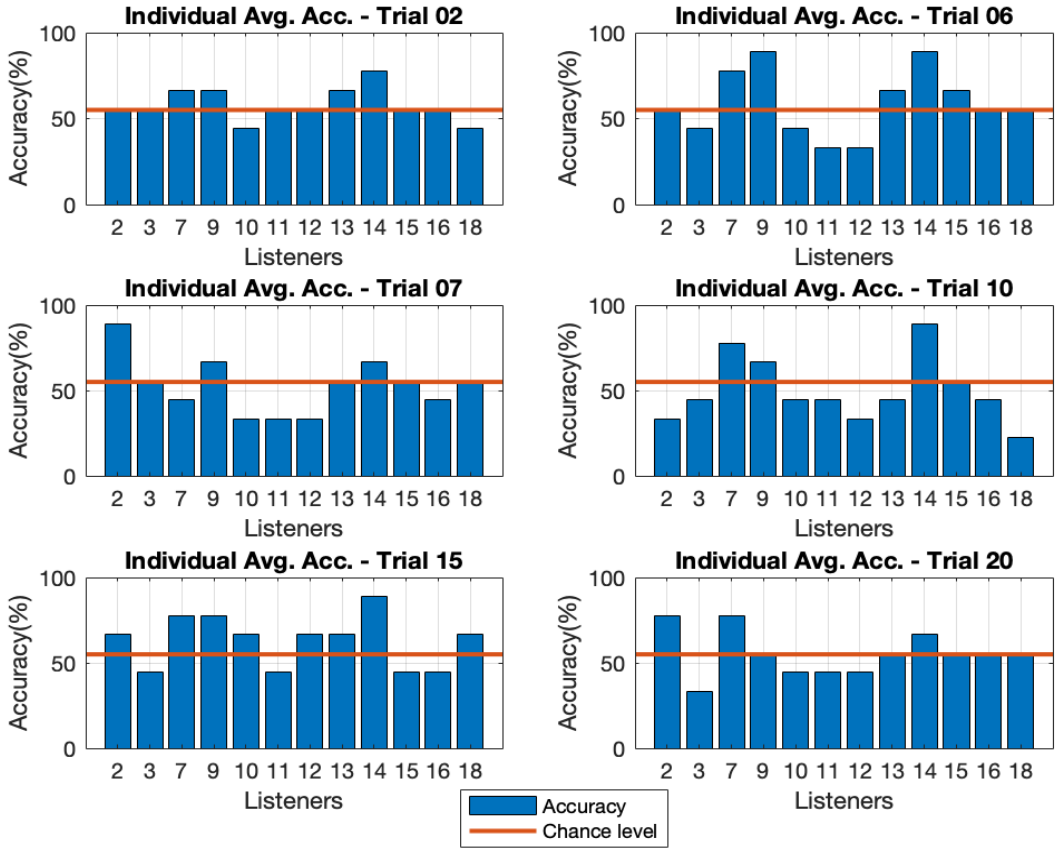


Figure 3.9: Average accuracy by listener with reduced data

certain listeners did not perform above chance level. This again signifies that the training was preferential to specific listeners, yet the increased overall accuracy across trials means the removal of certain trials was an overall benefit. Though the patterns are consistent there are only nine trials

per individual listener. The behavior of the network when more testing data in available would give a better overview to the performance for each individual listener, but it is limited by the data set used.

As was seen certain trials clearly perform better in the training than others. To access the overall contribution of each trial the combined resulting average accuracy was looked at. It was considered that as the trials with less preferable average loss trends were removed from the total accuracy it would increase. Figure 3.10 shows the resulting accuracy of including only the percentage of trials with the most negative average loss trend after all training. Each point includes a 90% confidence and the percentage for chance level performance is shown. Up to 75% of trials were excluded from the accuracy calculation which corresponds to including only 5 trials, as any less would be considered impractical. Each point corresponds to a single trial being removed from the average. The overall average performance can be seen to be increasing as the worst trials are removed. The confidence intervals consistently rise above chance level performance once around 15% of the trials are removed. Including this metric shows that the network also has variation in how it is training given which trial is used as the testing set. This could give an indication that the network trains better given certain conditions.

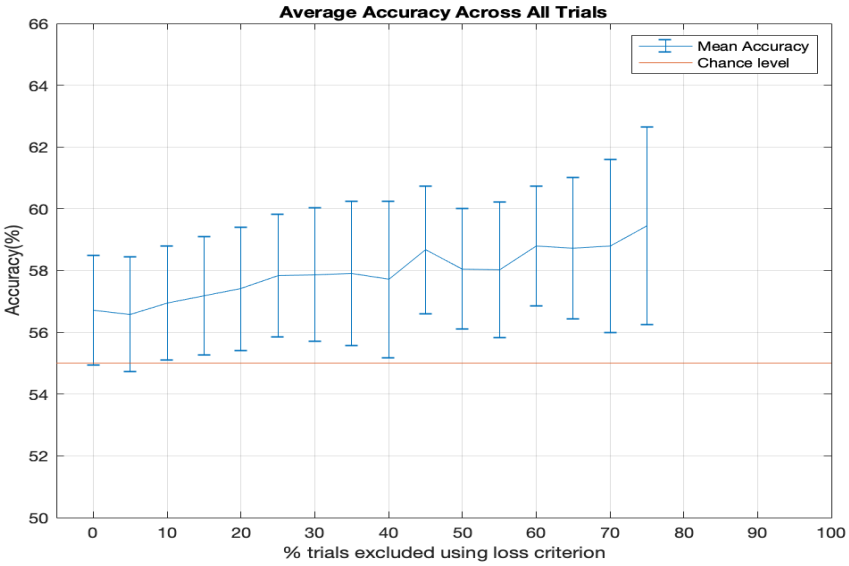


Figure 3.10: Overall accuracy as % of trials are excluded based on loss criterion

Chapter 4

Conclusion

In the implementation of a simple DNN for an AAD task it was shown that the method was robust to a significant reduction in available training data and an increase in subject variation within that data. The replication of the linear reconstruction filters was successful and provided reliability to the use of it as comparison. For these data sets it was shown that the anechoic condition training data did not perform as well on the simple DNN than the linear method. The overall decoding accuracy was shown to usually preform 20% worse than the linear method proposed in (Fuglsang et al.,2017). This is consistent with the difference in accuracy with the original data set that the DNN had. In terms of just the simple DNN performance on the new data set there were many similarities to the results of the original study. Mainly, the ability of the simple DNN model to effectively train on the new data set was clearly shown by the behavior of the average loss and accuracy functions throughout training. Decreasing trends in the average loss corresponding to increasing average accuracy trends. Once the listeners used where paired down the overall network accuracy for each trial improved to consistently reach above chance level performance. When separated by listener the accuracy on the newly introduced data showed distributions in accuracy similar to the previous work. This occurring despite the rather large reduction in testing samples per listener is promising for further expanded testing data.

Prior to this exploration into the networks each showed promising results, but the drastic difference in training data left it very difficult to compare the methods. For the same data set both models

were shown to output consistent results with their original papers. This gave legitimacy to both methods and gave a stronger criteria for comparison. Beyond just this the simple DNN model was shown to be able to perform with additional restrictions on available data. The fact that the DNN can be this robust gives good reason to further consider expansions on this model. It was interesting that so many listeners could not be successfully trained with the system, but there is clear success among the various listeners and testing trials considered.

4.1 Future Work

There are various directions that can be expanded on to better test many elements for both applications. First the performance of the simple DNN can be further explored. One specific metric is the performance of the network once certain trials are removed from the training data. Seeing the effects of decreasing the training data further would allow for another measurement of if the network is trained properly. Another metric would be the application of the training data from the low and high reverb data sets. Doing this would provide reasons for or against the use of the DNN in closer approximations if real-world environments. It would also act as another way to compare the network's performance with the linear method. Even further (Fuglesang et al.,2017) showed that reducing EEG channels needed to train the decoder was possible and may be a way to reduce the noise present in training for all methods. It can may reduce the computational complexity of the network.

There would also be a reasonable benefit in the expansion of the DNN architecture. Many different DNN models have already been proposed for AAD tasks. This simplicity of the DNN was a key feature, but there are also certain aspects that are not as clear for their application in smaller data sets. Further tests with alterations to the DNN could prove more appropriate for the proposed task. Finally, the implementation of this data set in other proposed AAD methods would also be another way to proceed in comparing model effectiveness. This would contribute to an overall goal of showing how many of these methods perform on the same data-set. The fact that this data set is rather small in comparison to many of the ones used to train the models it can also be a measure for how robust the model is. Despite the usefulness of this the data set is still not a measure for

the applicability of a model in a real-time system. There is still much progress that needs to be made in this regard. However, there is still a need for consistent and reliable comparison between the multitude of proposed state-of-the-art models.

References

[1] Sahar Akram, Jonathan Z. Simon, and Behtash Babadi. "Dynamic Estimation of the Auditory Temporal Response Function From MEG in Competing-Speaker Environments". In: *IEEE Transactions on Biomedical Engineering* 64.8 (Aug. 2017), pp. 1896–1905. ISSN: 0018-9294. DOI: 10.1109/TBME.2016.2628884. URL: http://ieeexplore.ieee.org/document/7744504/.

[2] Sahar Akram et al. "Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling". In: *NeuroImage* 124 (Jan. 2016), pp. 906–917. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2015.09.048. URL: https://www.isr.umd.edu/Labs/CSSL/simonlab/pubs/AkramEtAlNeuroImage2015.pdf%20http://linkinghub.elsevier.com/retrieve/pii/S1053811915008708%20http://www.sciencedirect.com/science/article/pii/S1053811915008708.

[3] S R Arnott and C Alain. "Effects of perceptual context on event-related brain potentials during auditory spatial attention". In: 39.5 (2002), pp. 625–632. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve%7B%5C&%7Ddb=PubMed%7B%5C&%7Ddopt=Citation%7B%5C&%7Dlist%7B%5C_%7Duids=12236329.

[4] Stephen R. Arnott and Claude Alain. "Effects of perceptual context on event-related brain potentials during auditory spatial attention". In: *Psychophysiology* 39.5 (Sept. 2002), S0048577202394149. ISSN: 00485772. DOI: 10.1017/S0048577202394149. URL: http://doi.wiley.com/10.1017/S0048577202394149.

[5] Ali Aroudi and Simon Doclo. "EEG-based auditory attention decoding using unprocessed binaural signals in reverberant and noisy conditions?" In: *2017 39th Annual International*

Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). Seogwipo: IEEE, July 2017, pp. 484–488. ISBN: 978-1-5090-2809-2. DOI: 10.1109/EMBC.2017.8036867. URL: https://ieeexplore.ieee.org/document/8036867/.

[6] Ali Aroudi et al. "Auditory attention decoding with EEG recordings using noisy acoustic reference signals". In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Shanghai: IEEE, Mar. 2016, pp. 694–698. ISBN: 978-1-4799-9988-0. DOI: 10.1109/ICASSP.2016.7471764. URL: http://ieeexplore.ieee.org/document/7471764/.

[7] Ali Aroudi et al. "Impact of Different Acoustic Components on EEG-Based Auditory Attention Decoding in Noisy and Reverberant Conditions". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 27.4 (Apr. 2019), pp. 652–663. ISSN: 1534-4320. DOI: 10.1109/TNSRE.2019.2903404. URL: https://ieeexplore.ieee.org/document/8662636/.

[8] Wouter Biesmans et al. "Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.5 (May 2017), pp. 402–412. ISSN: 15344320. DOI: 10.1109/TNSRE.2016.2571900. URL: http://ieeexplore.ieee.org/document/7478117/.

[9] A Bregman and P Ahad. *Demonstrations of auditory scene analysis: The perceptual organization of sound.Dem.* 1996.

[10] Michael P. Broderick et al. "Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech". In: *Current Biology* 28.5 (Mar. 2018), 803–809.e3. ISSN: 09609822. DOI: 10.1016/j.cub.2018.01.080. URL: https://linkinghub.elsevier.com/retrieve/pii/S0960982218301465.

[11] Michael P Broderick et al. "Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech". In: *Current Biology* 28.5 (Mar. 2018), 803–809.e3. ISSN: 09609822. DOI: 10.1016/j.cub.2018.01.080. URL: http://www.ncbi.nlm.nih.gov/pubmed/29478856%20http://linkinghub.elsevier.com/retrieve/pii/S0960982218301465.

[12] John F. Brugge et al. "Functional localization of auditory cortical fields of human: Click-train stimulation". In: *Hearing Research* 238.1-2 (Apr. 2008), pp. 12–24. ISSN: 03785955. DOI:

10.1016/j.heares.2007.11.012. URL: https://linkinghub.elsevier.com/retrieve/pii/S0378595507002730.

[13] E C Cherry. "Some experiments on the recognition of speech, with one and with two ears". In: *Journal of the Acoustical Society of America* 25.5 (1953), pp. 975–979.

[14] Alain de Cheveigné et al. "Decoding the auditory brain with canonical component analysis". In: *NeuroImage* 172 (May 2018), pp. 206–216. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2018.01.033. URL: https://linkinghub.elsevier.com/retrieve/pii/S1053811918300338.

[15] G. Ciccarelli et al. "Comparison of Two-Talker Attention Decoding from EEG with Nonlinear Neural Networks and Linear Methods". In: *Sci Rep* 9 (2019), p. 11538. URL: https://www.nature.com/articles/s41598-019-47795-0.

[16] Piers Dawes et al. "Hearing Loss and Cognition: The Role of Hearing Aids, Social Isolation and Depression". In: *PLOS ONE* 10.3 (Mar. 2015). Ed. by Blake Johnson, e0119616. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0119616. URL: https://dx.plos.org/10.1371/journal.pone.0119616.

[17] Giovanni M. Di Liberto, James A. O'Sullivan, and Edmund C. Lalor. "Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing". In: *Current Biology* 25.19 (Oct. 2015), pp. 2457–2465. ISSN: 09609822. DOI: 10.1016/j.cub.2015.08.030. URL: https://linkinghub.elsevier.com/retrieve/pii/S0960982215010015.

[18] K.V. Dijkstra et al. "Identifying the attended speaker using electrocorticographic (ECoG) signals". In: *Brain-Computer Interfaces* 2.4 (Oct. 2015), pp. 161–173. ISSN: 2326-263X. DOI: 10.1080/2326263X.2015.1063363. URL: http://www.tandfonline.com/doi/full/10.1080/2326263X.2015.1063363.

[19] N Ding and J Z Simon. "Emergence of neural encoding of auditory objects while listening to competing speakers". In: *Proceedings of the National Academy of Sciences* 109.29 (July 2012), pp. 11854–11859. ISSN: 0027-8424. DOI: 10.1073/pnas.1205381109. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1205381109.

[20] N Ding and J. Z. Simon. "Emergence of neural encoding of auditory objects while listening to competing speakers". In: *Proceedings of the National Academy of Sciences* 109.29 (July 2012), pp. 11854–11859. ISSN: 0027-8424. DOI: 10.1073/pnas.1205381109. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1205381109.

[21] Nai Ding and Jonathan Z. Simon. "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening". In: *Journal of Neurophysiology* 107.1 (Jan. 2012), pp. 78–89. ISSN: 0022-3077. DOI: 10.1152/jn.00297.2011. URL: http://www.physiology.org/doi/10.1152/jn.00297.2011%20http://jn.physiology.org/cgi/doi/10.1152/jn.00297.2011.

[22] J Driver. "A selective review of selective attention research from the past century". In: *British Journal of Psychology* 92.1 (2001), pp. 53–78.

[23] Fan-Gang Zeng et al. "Cochlear Implants: System Design, Integration, and Evaluation". In: *IEEE Reviews in Biomedical Engineering* 1 (2008), pp. 115–142. ISSN: 1937-3333. DOI: 10.1109/RBME.2008.2008250. URL: http://ieeexplore.ieee.org/document/4664429/.

[24] Ian C. Fiebelkorn, Yuri B. Saalmann, and Sabine Kastner. "Rhythmic Sampling within and between Objects despite Sustained Attention at a Cued Location". In: *Current Biology* 23.24 (Dec. 2013), pp. 2553–2558. ISSN: 09609822. DOI: 10.1016/j.cub.2013.10.063. URL: https://linkinghub.elsevier.com/retrieve/pii/S0960982213013389.

[25] Søren Fuglsang, Torsten Dau, and Jens Hjortkjær. "Noise-robust cortical tracking of attended speech in real-world acoustic scenes". In: *NeuroImage* 156 (Aug. 2017), pp. 435–444. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2017.04.026. URL: http://www.sciencedirect.com/science/article/pii/S105381191730318X?via%7B%5C%%7D3Dihub%20http://linkinghub.elsevier.com/retrieve/pii/S105381191730318X.

[26] M Giard et al. "Neurophysiology Mechanisms of Auditory Selective Attention in Humans". In: *Frontiers in bioscience* 5 (2000), pp. 84–94. URL: http://bio-mirror.im.ac.cn/mirrors/bioscience/2000/v5/d/giard/giard.pdf.

[27] T D Griffiths and J D Warren. "What is an auditory object?" In: *Nature neurosc.reviews* 5.11 (2004), pp. 887–892. URL: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?

cmd=Retrieve%7B%5C&%7Ddb=PubMed%7B%5C&%7Ddopt=Citation%7B%5C&%7Dlist%7B%5C_%7Duids=15496866.

[28] Cong Han et al. "Speaker-independent auditory attention decoding without access to clean speech sources". In: *Science Advances* 5.5 (May 2019), eaav6134. ISSN: 2375-2548. DOI: 10.1126/sciadv.aav6134. URL: https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.aav6134.

[29] G Hinton et al. "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups". In: *Signal Processing Magazine, IEEE* 29.6 (2012), pp. 82–97. DOI: 10.1109/MSP.2012.2205597.

[30] Colin Humphries, Einat Liebenthal, and Jeffrey R. Binder. "Tonotopic organization of human auditory cortex". In: *NeuroImage* 50.3 (Apr. 2010), pp. 1202–1211. ISSN: 10538119. DOI: 10.1016/j.neuroimage.2010.01.046. URL: https://linkinghub.elsevier.com/retrieve/pii/S1053811910000686.

[31] S Kähkönen. "Effects of Haloperidol on Selective Attention A Combined Whole-Head MEG and High-Resolution EEG Study". In: *Neuropsychopharmacology* 25.4 (Oct. 2001), pp. 498–504. ISSN: 0893133X. DOI: 10.1016/S0893-133X(01)00255-X. URL: http://www.nature.com/doifinder/10.1016/S0893-133X(01)00255-X.

[32] Ozlem Kalinli and Shrikanth Narayanan. "A top-down auditory attention model for learning task dependent influences on prominence detection in speech". In: *2008 IEEE International Conference on Acoustics, Speech and Signal Processing, VOLS 1-12*. International Conference on Acoustics Speech and Signal Processing (ICASSP). 2008, pp. 3981–3984.

[33] Emine Merve Kaya and Mounya Elhilali. "Investigating bottom-up auditory attention". In: *Frontiers in Human Neuroscience* 8 (May 2014), p. 327. ISSN: 1662-5161. DOI: 10.3389/fnhum.2014.00327. URL: http://journal.frontiersin.org/article/10.3389/fnhum.2014.00327/abstract%20http://www.ncbi.nlm.nih.gov/pubmed/24904367%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4034154.

[34] Edmund C Lalor and John J Foxe. "Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution". In: *European Journal of Neuroscience* 31.1 (2010), pp. 189–193. DOI: 10.1111/j.1460-9568.2009.07055.x.

[35] David Looney et al. "The In-the-Ear Recording Concept: User-Centered and Wearable Brain Monitoring". In: *IEEE Pulse* 3.6 (Nov. 2012), pp. 32–42. ISSN: 2154-2287. DOI: 10.1109/MPUL.2012.2216717. URL: http://ieeexplore.ieee.org/document/6378569/.

[36] D Looney et al. "An in-the-ear platform for recording electroencephalogram". In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Boston, MA: IEEE, Aug. 2011, pp. 6882–6885. ISBN: 978-1-4577-1589-1. DOI: 10.1109/IEMBS.2011.6091733. URL: http://ieeexplore.ieee.org/document/6091733/.

[37] Fernando Lopes da Silva. "EEG and MEG: Relevance to Neuroscience". In: *Neuron* 80.5 (Dec. 2013), pp. 1112–1128. ISSN: 08966273. DOI: 10.1016/j.neuron.2013.10.017. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627313009203.

[38] Huan Luo and David Poeppel. "Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex". In: *Neuron* 54.6 (June 2007), pp. 1001–1010. ISSN: 08966273. DOI: 10.1016/j.neuron.2007.06.004. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627307004138.

[39] David J. Mener et al. "Hearing Loss and Depression in Older Adults". In: *Journal of the American Geriatrics Society* 61.9 (Sept. 2013), pp. 1627–1629. ISSN: 00028614. DOI: 10.1111/jgs.12429. URL: http://doi.wiley.com/10.1111/jgs.12429.

[40] Nima Mesgarani and Edward F Chang. "Selective cortical representation of attended speaker in multi-talker speech perception". In: *Nature* 485.7397 (May 2012), pp. 233–236. ISSN: 0028-0836. DOI: 10.1038/nature11020. URL: http://www.nature.com/articles/nature11020.

[41] Paul Mick, Ichiro Kawachi, and Frank R. Lin. "The Association between Hearing Loss and Social Isolation in Older Adults". In: *Otolaryngology–Head and Neck Surgery* 150.3 (Mar. 2014), pp. 378–384. ISSN: 0194-5998. DOI: 10.1177/0194599813518021. URL: http://journals.sagepub.com/doi/10.1177/0194599813518021.

[42] Mohammad Jalilpour Monesi et al. "An LSTM Based Architecture to Relate Speech Stimulus to EEG". In: (Feb. 2020). arXiv: 2002.10988. URL: http://arxiv.org/abs/2002.10988.

[43] J A Müller et al. "Influence of auditory attention on sentence recognition captured by the neural phase". In: *European Journal of Neuroscience* 51 (2020), pp. 1305–1314. DOI: 10.1111/ejn.13896.

[44] R Naatanen. "The Role of Attention in Auditory Information-Processing as Revealed by Event-Related Potentials and Other Brain Measures of Cognitive Function". In: 13.2 (1990), pp. 201–232. URL: %7B%5C%%7D3CGo%20to.

[45] Rochelle S. Newman. "The Cocktail Party Effect in Infants Revisited: Listening to One's Name in Noise." In: *Developmental Psychology* 41.2 (2005), pp. 352–362. ISSN: 1939-0599. DOI: 10.1037/0012-1649.41.2.352. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0012-1649.41.2.352.

[46] James A O'Sullivan et al. "Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG". In: *Cerebral Cortex* 25.7 (July 2015), pp. 1697–1706. ISSN: 1460-2199. DOI: 10.1093/cercor/bht355. URL: http://cercor.oxfordjournals.org/content/early/2014/01/14/cercor.bht355.abstract%20https://academic.oup.com/cercor/article-lookup/doi/10.1093/cercor/bht355.

[47] C Pantev et al. "Specific tonotopic organizations of different areas of the human auditory cortex revealed by simultaneous magnetic and electric recordings". In: *Electroencephalography and Clinical Neurophysiology* 94.1 (Jan. 1995), pp. 26–40. ISSN: 00134694. DOI: 10.1016/0013-4694(94)00209-4. URL: https://linkinghub.elsevier.com/retrieve/pii/0013469494002094.

[48] Dana J. Plude, Jim T. Enns, and Darlene Brodeur. "The development of selective attention: A life-span overview". In: *Acta Psychologica* 86.2-3 (Aug. 1994), pp. 227–272. ISSN: 00016918. DOI: 10.1016/0001-6918(94)90004-3. URL: https://linkinghub.elsevier.com/retrieve/pii/0001691894900043.

[49] M. Sawan et al. "Wireless Recording Systems: From Noninvasive EEG-NIRS to Invasive EEG Devices". In: *IEEE Transactions on Biomedical Circuits and Systems* 7.2 (Apr. 2013), pp. 186–195. ISSN: 1932-4545. DOI: 10.1109/TBCAS.2013.2255595. URL: http://ieeexplore.ieee.org/document/6508907/.

[50] Christoph E. Schreiner, Heather L. Read, and Mitchell L. Sutter. "Modular Organization of Frequency Integration in Primary Auditory Cortex". In: *Annual Review of Neuroscience* 23.1 (Mar. 2000), pp. 501–529. ISSN: 0147-006X. DOI: 10.1146/annurev.neuro.23.1.501. URL: http://www.annualreviews.org/doi/10.1146/annurev.neuro.23.1.501.

[51] Shihab A. Shamma, Mounya Elhilali, and Christophe Micheyl. "Temporal coherence and attention in auditory scene analysis". In: *Trends in neurosciences* 34.3 (Mar. 2011), pp. 114–23. ISSN: 1878-108X. DOI: 10.1016/j.tins.2010.11.002. URL: http://linkinghub.elsevier.com/retrieve/pii/S0166223610001670%20http://www.ncbi.nlm.nih.gov/pubmed/21196054%20http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3073558.

[52] Barbara G. Shinn-Cunningham. "Object-based auditory and visual attention". In: *Trends in Cognitive Sciences* 12.5 (May 2008), pp. 182–186. ISSN: 13646613. DOI: 10.1016/j.tics.2008.02.003. URL: https://linkinghub.elsevier.com/retrieve/pii/S1364661308000600.

[53] E Sussman, W Ritter, and H G Vaughan Jr. "An investigation of the auditory streaming effect using event-related brain potentials". In: *Psychophysiology* 36.1 (Jan. 1999), pp. 22–34.

[54] Tobias Taillez, Birger Kollmeier, and Bernd T Meyer. "Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech". In: *European Journal of Neuroscience* 51.5 (Mar. 2020), pp. 1234–1241. ISSN: 0953-816X. DOI: 10.1111/ejn.13790. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/ejn.13790.

[55] USVA. *Annual Benefits Report Fiscal Year 2017*. Tech. rep. US Department of Veterans Affairs, Veterans Benefits Administration, 2017.

[56] S Vandecappelle et al. "EEG-based detection of the attended speaker and the locus of auditory attention with convolutional neural networks". Luven, Belgium, 2020. URL: https://doi.org/10.1101/475673.

[57] Michael Vongphoe and Fan-Gang Zeng. "Speaker recognition with temporal cues in acoustic and electric hearing". In: *The Journal of the Acoustical Society of America* 118.2 (Aug. 2005), pp. 1055–1061. ISSN: 0001-4966. DOI: 10.1121/1.1944507. URL: http://asa.scitation.org/doi/10.1121/1.1944507.

[58] Blake S Wilson et al. "Global hearing health care: new findings and perspectives". In: *The Lancet* 390.10111 (Dec. 2017), pp. 2503–2515. ISSN: 01406736. DOI: 10.1016/S0140-6736(17)31073-5. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673617310735.

[59] Daniel Wong et al. “A comparison of regularization methods in forward and backward models for auditory attention decoding”. In: *Frontiers in Neuroscience* 12.AUG (Aug. 2018), p. 531. ISSN: 1662453X. DOI: 10.3389/fnins.2018.00531. URL: https://www.frontiersin.org/article/10.3389/fnins.2018.00531/full.

[60] Mengni Zhou et al. “Epileptic Seizure Detection Based on EEG Signals and CNN”. In: *Frontiers in Neuroinformatics* 12 (Dec. 2018), p. 95. ISSN: 1662-5196. DOI: 10.3389/fninf.2018.00095. URL: https://www.frontiersin.org/article/10.3389/fninf.2018.00095/full.

[61] Elana M Zion Golumbic, David Poeppel, and Charles E. Schroeder. “Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective”. In: *Brain and Language* 122.3 (Sept. 2012), pp. 151–161. ISSN: 0093934X. DOI: 10.1016/j.bandl.2011.12.010. URL: https://linkinghub.elsevier.com/retrieve/pii/S0093934X11001982.

[62] Benedikt Zoefel and Rufin VanRullen. “EEG oscillations entrain their phase to high-level features of speech sound”. In: *NeuroImage* 124 (Jan. 2016), pp. 16–23. ISSN: 10959572. DOI: 10.1016/j.neuroimage.2015.08.054. URL: https://www.sciencedirect.com/science/article/pii/S1053811915007739.

Jack C. Magann

jmagann1@jhu.edu 301-502-1896

LinkedIn: www.linkedin.com/in/jackmagann

Education

Johns Hopkins University Expected Graduation - May 2020
M.S.E Electrical Engineering GPA: 3.33
Lab for Computational Audio Perception directed by Mounya Elhilali

Relevant coursework: Audio Signal Processing, Machine Learning for Signal Processing, Random Signal Analysis, Compressed Sensing and Sparse Recovery, Information Extraction

University of Miami May 2018
B.S. Music Engineering GPA: 3.77
Minors - Electrical Engineering, Computer Engineering

Relevant coursework: Digital Signal Processing, Engineering Acoustics, Transducer Theory, Audio Plug-in Design, Synth Design, Microprocessors, Computational Psychoacoustics

Experience

Assistant Signal Processing Engineer Summer 2018
University of Miami - FORE Lab

- Programmed audio effects for a mobile app being developed to help veterans with leg prosthetics learn to walk correctly through the distortion of music

Auditory Localization Research Assistant - REU Program Summer 2017
Michigan State University

- Conducted psychoacoustic research on the localization of Sine Tones by Human listeners in a room
- Focused on the analysis of inter-aural queues to distinguish front-back confusions
- Research submitted and accepted to be presented at the Acoustical Society of America (ASA) 174th meeting (Dec 2017).

Head Recording Engineer Aug. 2016 - May 2018
University of Miami - Recording Services

- Recorded performances in the University's Concert halls and provided training to new engineers

Head Live Sound Engineer Sep. - Dec. 2015/2016
University of Miami - Band of the Hour

- Lead a team to provide live sound enhancement to the marching band during all performances

Technical Skills

- Knowledge of Digital Signal Processing and Random Signal Analysis
- Experience in auditory perception and attention decoding research
- Experience using TensorFlow for machine learning algorithms focused on audio signal processing
- Experience implementing machine learning algorithms for signal processing including GMMs, HMMs, SVM, etc...
- Advanced MATLAB skills, Intermediate C++ and python 2/3 skills, basic Assembly knowledge
- Audio Plug-In and Synth Design in C++
- Experience designing audio analog circuits and hardware
- Experience in mixing and music production in Protools 10, Ableton Live 9 and Logic X
- Experience using PSPICE, Circuit Maker, ExpressPCB, Microsoft office (Excel, Word, etc. . .)
- 13 years experience as a classical and contemporary trumpet player

Leadership/Volunteering

- Volunteer at Seasons Hospice Care
- Social Chair for the Graduate Representative Organization at Johns Hopkins University
- Facilitator for Safe Zone Training at Johns Hopkins University
- 3 years on Elective Board for the University of Miami Running Club
- 1 year on Elective Board for University of Miami Association of Computing Machinery
- 3 years playing for the University of Miami Pep Band
- Chair of the Graduate Queer-Straight Alliance at Johns Hopkins University

Teaching Experience

Teaching Assistant Spring 2019/2020
Johns Hopkins University - Signals and Systems

Publications

Eric J. Macaulay, Jack C. Magann, and William M. Hartmann. The effect of listener motion on localization of tones in a room. (2017). The Journal of the Acoustical Society of America 142, 2676